Two-Layers re-Ranking Approach based on Contextual Information for Visual Concepts Detection in Videos

Abdelkader Hamadi, Georges Quénot, Philippe Mulhem

Two-layers re-ranking approach based on contextual information for visual concepts detection in videos

Abdelkader HAMADI Georges QUÉNOT Philippe MULHEM
{Abdelkader.Hamadi, Georges.Quenot, Philippe.Mulhem}@imag.fr
UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217, Grenoble, F-38041, France

Abstract

Context helps to understand the meaning of a word and allows the disambiguation of polysemic terms. Many researches took advantage of this notion in information retrieval. For concept-based video indexing and retrieval, this idea seems a priori *valid. One of the major problems is then to provide a definition of the context and to choose the most appropriate methods for using it. Two kinds of contexts were exploited in the past to improve concepts detection: in some works, inter-concepts relations are used as semantic context, where other approaches use the temporal features of videos to improve concepts detection. Results of these works showed that the "temporal" and the "semantic" contexts can improve concept detection. In this work we use the semantic context through an ontology and exploit the efficiency of the temporal context in a "two-layers" re-ranking approach. Experiments conducted on TRECVID 2010 data show that the proposed approach always improves over initial results obtained using either MSVM or KNN classifiers or their late fusion, achieving relative gains between* 9% *and* 33% *of the MAP measure.*

keywords : Semantic video indexing, Conceptual context, Inter-concepts relationships, Temporal context, Ontology, TRECVID.

1. Introduction

This paper deals with concept detection in videos. The main idea tackled here is that the occurrence of a concept in a video depends on specific contexts, in addition to the usual learning frameworks used for automatic video concept detection. Such use of context proved its interest [5, 9, 12]. We define here a context for video concept detection as any information that might improve an initial concept detection, by re-weighting (and therefore *re-ranking*) the concepts detected. Two questions arise here: a) which contexts might be effective? and b) how several contexts might be combined effectively? To our knowledge, such combinations have not been used in detail in the past. To answer these questions, we propose: i) one context based on a mixed ontology/corpus that exhibits relevant families of concepts; ii) one context that takes benefit from the temporal sequence of shots in videos, inspired from [9, 7]; and iii) several two-layers re-ranking approaches that combine these contexts. The contexts i) and ii) above are very different, and we believe that combining them has best chances to improve each context taken separately.

The remainder of this paper is organized as follows. Section 2 focuses on related works. Section 3 describes our proposal on context-based concept detection and two-layers re-rankings. Section 4 discusses the experiments on the TRECVID 2010 corpus and results achieved using two state of the art learning frameworks. We conclude in the last section.

2. Related Works

Several researchers exploit inter-concepts relations and temporal context to improve concept detection systems' performance. Some works using concepts relations are based on ontologies. [10] uses the "ancestors" of concepts in an ontology to improve initial detection scores of the descendants concepts. Other works model the relations between concepts based on datasets [8, 12, 2, 3, 4]. In [3] a probabilistic framework of

multijects and the multinet was proposed. The authors used the multinet to model context in terms of concepts co-occurrence. All of these works showed that the use of inter-concepts relations is beneficial for the detection of concepts. However, few works combine information based on dataset and information coming from a human expert. In the other hand, Safadi and Quénot [7] propose an effective approach based on local homogeneity of videos exploiting the temporal dependency between shots to improve concept detection.

In this work we use the semantic context by mixing the use of an ontology and a corpus; in the other hand, we exploit the efficiency of the temporal context in a "two-layers" re-ranking approach, we believe that combining the two contexts has better chances to improve each context taken separately.

3 Proposed Approaches

3.1. Ontology based approaches

Using relations between concepts extracted from data is a good idea, but a human expert input may also be crucial. Such human input may take the form of an ontology. The main question that we can ask is: "Could a human post process of an existing ontology improve video concept detection?". The use of a predefined ontology can be criticized for several reasons: giving equal weights to two arcs connecting two pairs of concepts having different similarity degrees. On the other hand, an ontology does not consider the data sets, which does not make it suitable for all types of data, leading to the need to adapt it to the data. To overcome these drawbacks, i) we use an ontology to determine interrelated concepts, ii) we do not use paths connecting the concepts to calculate the distance between concepts but weights are calculated based on data sets. This permits to use simultaneously the ontology and data information. Inspired by [12] we use correlation coefficients as weights:

$$Corr(c_i, c_j) = \frac{\sum_{k=1}^{M} (S_{c_i}^k - \bar{S_{c_i}})(S_{c_j}^k - \bar{S_{c_j}})}{\sqrt{\sum_{k=1}^{M} (S_{c_i}^k - \bar{S_{c_i}})^2} \sqrt{\sum_{k=1}^{M} (S_{c_j}^k - \bar{S_{c_j}})^2}}$$

M: data size(number of shots)
$S_{c_i}^k = 1$ if c_i is present in the k^{th} shot, -1 otherwise
$\bar{S_{c_i}}$: = mean of $S_{c_i}^k$ values $= (\sum_{k=1}^{M} S_{c_i}^k)/M$

Let us note sc_{c_i} the initial detection score of the concept C_i achieved by a learner, and sc'_{c_i} the new detection score of C_i after the re-ranking step. We propose two approaches to determine which concepts to be combined:

1. "Ancestors or Descendants": For a concept C_i, combining only concepts that are ancestors or descendants of C_i

$$sc'_{c_i} = sc_{c_i} + \sum_{c_j \in ont(c_i)} Corr(c_i, c_j).sc_{c_j}$$

where $ont(c_i) = \{c_j |\ c_j$ is an ancestor or a descendant of c_i in the ontology$\}$.

2. "Concept family": For a concept C_i, a human expert selects the concepts related to concept C_i. Based on our experiments, we found that concepts co-occurrences show that for a concept C_i the concepts that are ancestors, descendants in the ontology, as well as other concepts that share a same ancestor of C_i co-occur usually with C_i. Other experiments confirmed that not all of these concepts help to detect C_i. From these observations, we propose that a human expert selects for each concept C_i a set of concepts representing what we will call "concept family": the latter is a set of concepts related semantically to the concept C_i. For example for the concept "Car" family contains all the "transportation" concepts, and the concept "Cat" will have all "animals" concepts in its family. Assuming that our ontology is a tree, a human expert divides it into a forest where each tree represents a hierarchy of concepts with common sense. Then, an initial family of a concept C_i is a set containing all descendants of each

ancestor of C_i (the whole tree to which it belongs). After selecting the initial family of concepts and based on a development corpus, we eliminate concepts that do not help to improve the detection of C_i. We calculate the new detection score of C_i by the following formula:

$$sc'_{c_i} = sc_{c_i} + \sum_{c_j \in F(c_i)} Corr(c_i, c_j).sc_{c_j}$$

$F(c_i) = \{$all concepts that we (human expert) consider as related to the concept $c_i\}$.

3.2. Temporal context approach

In addition to audio, a video has a feature that makes it different from an image: the temporal aspect. Ignoring such major characteristic in concept detection may lose relevant information. In fact, unlike an image video shots are linked and to understand its content we need information contained in a set of successive shots. [7] used the temporal context notion, exploiting the concept detection scores in the neighboring shots, leading to a very significant improvement. This result can be explained according to content dependency between locally homogeneous successive shots. We propose to use this idea in our work, and we extend our research of a concept on a window of size $2w + 1$ consecutive shots (w shots before and w shots after the current shot) . The new detection score for a concept C_i in the shot j is given by the following formula:

$$sc'^{(w)}_{S^j_{c_i}} = sc_{S^j_{c_i}} + \sum_{k=-w}^{w} sc_{S^{j+k}_{c_i}}$$

where $sc_{S^t_{c_i}}$ is the detection score for the concept C_i in the shot t and w the window size.

3.3. Two-layers re-ranking approach

The semantic context (inter-concepts relations) and the temporal context are different and important for visual concepts detection, based on this assumption we propose to combine the two approaches: "concepts family" and "temporal context". We propose a "two layers " re-ranking approach which consists of applying a first re-ranking method, then uses a second re-ranking based on the results of the first step. Because it is difficult to foresee *a priori* which combination is the best, we propose three possible combinations:

1. fusion: merging results of both approaches by averaging the concept detection scores for each shot:

   $$sc''_{S^j_{c_i}} = (scf'_{S^j_{c_i}} + sct'_{S^j_{c_i}})/2$$

   sc'': the new detection score of C_i in shot j. scf', sct': detection scores using respectively "concept family" and the "temporal context" approaches;
2. application of the "temporal re-ranking" on the results of "concepts family" approach;
3. application of the"concepts family" on the results of the "temporal context".

4. Experiments & Results

We tested and evaluated the above described approaches in the context of the TRECVID 2010 semantic indexing task. The re-ranking step is related to concept detection scores provided by individual concept detectors. Safadi and Quénot used Multi-SVMs(MSVM)[6] and achieved good results. In the other hand, [11] showed that the use of KNN as concept detector is a good alternative. In fact, because of their good results, we chose to use as supervised classifiers MSVM and a variant of KNN optimized separately for each concept (KNNC). As input of these learners, five descriptors were extracted from each shot as a representative feature vector. We tested the approaches with color, texture and interest points (SIFT) descriptors for the visual modality, MFCC based descriptors for the audio modality and combinations of them. These descriptors were extracted by LIG[2] and GIPSA[3] teams from the IRIM[4] group.

[2]http://www.liglab.fr/
[3]http://www.gipsa-lab.inpg.fr/
[4]http://mrim.imag.fr/irim/

Here are details about the individual descriptors used:

- LIG/$h3d64$ ($hist$): normalized RGB Histogram 4×4×4 (64-dim):
- LIG/$gab40$ (gab): normalized Gabor transform, 8 orientations × 5 scales (40-dim);
- LIG/$hg104$ (hg): early fusion (concatenation) of LIG/$h3d64$ and LIG/gab40 (104-dim);
- LIG/$opp\_sift\_har\_1000$ ($sift$): bag of words, opponent sift with Harris-Laplace detector (1000-dim);
- GIPSA/$AudioSpectro.b28$ ($audio$): spectral profile in 28 bands on a Mel scale.

We could use here many other descriptors. Because we focused on the re-ranking step, we made these simple choices, which are not the best to have a good system, but to test the robustness of our approaches we made a late fusion of results obtained by 40 single descriptors which gave a quite good system in terms of MAP.

We made a comparison of our proposal with the "boosting and confusion factors" approach detailed in [10] where an ontology and relations of the type "C_i excludes C_j" were used for re-ranking.

Our evaluation was conducted on TRECVID 2010 data set. We ran our experiments on the development set split into two parts: one for training and the other for evaluation ("1-fold cross-validation"). The annotations were provided by the TRECVID 2010 collaborative annotation organized by LIG and LIF [1]. We use a lexicon containing 130 concepts. The ontology used in our experiments was built based on a set of inter-concepts relations of the type "C_1 implies C_2" as follows: if C_1 implies C_2 then C_2 is an ancestor of C_1. The performance was measured by the Mean Average Precision (MAP) computed on the 130 concepts. We fixed $w = 5$ in our experiments for the size of the temporal window surrounding a shot. We made the following experiments:

1. after running MSVM and KNNC for each concept using the features cited above, we applied re-rankings;
2. we merged results obtained by individual concepts detectors using the five single descriptors (*fusion_desc*), we applied then re-rankings;
3. we made a late fusion of MSVM and KNNC scores for each descriptor even on $fusion\_desc$, we applied then re-rankings;
4. to test the robustness of our approach we applied re-rankings on the results of a system with a good performance in terms of MAP. We applied then re-rankings on *Quaero_fusion* scores which are obtained by applying a late fusion of 40 descriptors (textures, visual, audio, sift using MSVM as classifier). The value of MAP measured on these scores exceed 0.14. Note that in TRECVID 2011 the best system got about 0.2 as MAP value.

Because of a lack of space, we do not show in what follows all the details about the results obtained by applying re-rankings on the scores of each leaner (KNNC, MSVM) separately.

4.1. Results using ontology and temporal context approaches

Table 1 presents a comparison between the results obtained by applying our proposed approaches, the "Boosting" and "Confusion" factors methods. Re-rankings were applied on the results of the late fusion of KNNC and MSVM scores. $ConfusionFactor$ always deteriorates the results while $BoostingFactor$ performs better using single descriptors, but the improvement is less when the initial system is more efficient. We can see that the three approaches improve the intial results whatever kind of features used and perform better than $BoostingFactor$ whether for single descriptors or more efficient systems, but "concept Family"(conFamily) gets better results than "Ancestors or descendants" and the "temporal context" approach(*Temp*) gives the overall

		State of the art		Our approaches		
	$initial$	$BoostingF$	$ConfusionF$	$AncOrdesc$	$ConFamily$	Temporal Context
$hist$	0.0343	0.0345(+0.58)	0.0341(-0.58)	0.0347 (+1.17)	0.0356 (+3.79)	0.0398 **(+16.03)**
gab	0.0307	0.0309(+0.65)	0.0306(-0.32)	0.0311 (+1.30)	0.0315 (+2.60)	0.0337 **(+9.77)**
hg	0.0548	0.0549(+0.18)	0.0546(-0.36)	0.0550 (+0.36)	0.0559(+2.01)	0.0608**(+10.95)**
$sift$	0.0698	0.0710(+1.72)	0.0696(-0.28)	0.0711 (+1.86)	0.0725 (+3.87)	0.0782 **(+12.03)**
$audio$	0.0136	0.0138(+1.47)	0.0135(-0.73)	0.0142(+4.41)	0.0146 (+7.35)	0.0157 **(+15.44)**
$fusion\_desc$	0.0832	0.0844(+1.44)	0.0827 (-0.60)	0.0841 (+1.08)	0.0856 (+2.88)	0.0925 **(+11.18)**
$Quaero\_fusion$	0.1428	**0.1447**(+1.33)	**0.1419**(-0.63)	**0.1457**(+2.03)	**0.1478** (+3.50)	**0.1561(+9.31)**

Table 1. Results (MAP (gain %)) using different re-ranking approaches

	$initial$	$2layers\_Fusion$	Temp → ConFamily	ConFamily → Temp
$hist$	0.0343	0.0399 (+16.33)	0.0419 (+22.16)	0.0421 (**+22.74**)
gab	0.0307	0.0342 (+11.40)	0.0353 (+14.98)	0.0354 (**+15.31**)
hg	0.0548	0.0609 (+11.13)	0.0631 (+15.14)	0.0630 (**+14.96**)
$sift$	0.0698	0.0789 (+13.04)	0.0818 (+17.19)	0.0831 (**+19.05**)
$audio$	0.0136	0.0159 (+16.91)	0.0169 (+24.26)	0.0173 (**+27.20**)
$fusion\_desc$	0.0832	0.0944 (+13.46)	0.0976 (+17.31)	**0.0986** (**+18.51**)
$Quaero\_fusion$	0.1428	**0.1563** (+9.45)	**0.1577** (+10.43)	**0.1589** (**+11.27**)

Table 2. Results (MAP (gain %)) using two-layers re-ranking approach

best results. *ConFamily* achieves a gain between +2.01% and +7.35%. The best value of MAP is 0.1478 achieved when using $Quaero\_fusion$. For *Temp* the gain ranges between +9.77% and +16.03%. The best value of MAP is 0.1561 obtained when using *Quaero_fusion*. When applying re-rankings on the scores of each classifier separately the improvement is between +2.92% and +7.88% using *ConFamily* on KNNC scores and between +0.88% and +7.02% using *ConFamily* on MSVM scores. Regarding *Temp* the improvement achieved by using MSVM scores is between +9.72% and +21.05% while it ranges from +10.92% to +23.64% using KNNC scores.

We can see on table 1 that *Temp* improves concepts detection better than *ConFamily*; however, the MAP values of the two approaches are not intrinsically comparable. In fact, *Temp* uses the detection scores of not only the shot to index but also the scores of the neighbouring shots, while *ConFamily* uses the development corpus annotations and only the scores of the shot to index. We can explain the difference between the performances of both approaches by the fact that *Temp* merges scores obtained by the same learner while *ConFamily* makes a fusion of scores obtained by different classifiers trained independently, fact wich leads to a score normalization problem. The two approaches capture two different kinds of information; this is why we expect even better results when combining them.

4.2. Results using two-layers re-ranking

Table 2 presents the "two-layers" re-ranking results using the late fusion of KNNC and MSVM scores. This method improves the results whatever kind of features used; however, *2layers_Fusion* is less efficient compared to *Temp*. The two other approaches give better performances than *Temp*. The best results are achieved when applying the *Temp* on the results of *ConFamily* results. In such a case the improvement ranges between +14.96% and +27.20% and the best value of MAP is 0.1589 achieved by using *Quaero_fusion*. The gain when using MSVM scores ranges from +12.37% to +33.34%. The improvement when using KNNC results is between

+14.62% and +32.06%.

To summarize, the system performance when applying re-ranking on MSVM scores is better than when using KNNC results; however, the gain of re-ranking KNNC scores is higher. Moreover, the best performances are obtained when applying re-ranking on the results of the late fusion. Nevertheless, the gain is lower than when using the results of each classifier separately, we believe that this is because ther better the original results are, the harder it is to improve them.

5. Conclusion & future works

We proposed an approach exploiting inter-concepts relations, "concepts family", and we combine it in a "two layers" re-ranking approach with a temporal context based re-ranking method. The results showed that this combination gives the best performance compared to using each method separately, we obtained the best improvement by applying the "temporal" re-ranking on the results of the "concepts family" approach. Experiments conducted on TRECVID 2010 data show that the proposed approach always improves over initial results obtained using either MSVM or KNN classifiers or their late fusion, achieving relative gains between 9% and 33% of the MAP measure.

A MAP value of about 0.10-0.15 may seem quite low but a indexing system with such performance is already quite usable especially considering that this is an average of the precision at al levels of recall and that the precision at the top of the sorted list is much higher.

Concerning future works, we shall try to avoid the penality of errors propagation in the re-ranking stage. This could be achieved by exploiting inter-concepts relations in a none-re-ranking approach, for example by using relations as an input of the learning algorithms or during the learning step. Adding to that, refining the definition of the context and scores normalization are interesting prospects.

6. Acknowledgements

This work was partly realized as part of the Quaero Programme funded by OSEO, French State agency for innovation.

References

[1] Stéphane Ayache and Georges Quénot. Video corpus annotation using active learning. In *Proceedings of the IR research*, ECIR'08, pages 187–198, Berlin, Heidelberg, 2008. Springer-Verlag.

[2] Yusuf Aytar, O. Bilal Orhan, and Mubarak Shah. Improving semantic concept detection and retrieval using contextual estimates. In *ICME*, pages 536–539, 2007.

[3] M.R. Naphade and T.S. Huang. Detecting semantic concepts using context and audiovisual features. In *Detection and Recognition of Events in Video, 2001. Proceedings. IEEE Workshop on*, pages 92 –98, 2001.

[4] M.R. Naphade and T.S. Huang. Recognizing high-level audio-visual concepts using context. In *Image Processing, 2001. Proceedings. 2001 International Conference on*, volume 3, pages 46 –49 vol.3, 2001.

[5] Yu Qiu, Genliang Guan, Zhiyong Wang, and Dagan Feng. Improving news video annotation with semantic context. In *Proceedings of the 2010 International Conference on Digital Image Computing: Techniques and Applications*, DICTA '10, pages 214–219, Washington, DC, USA, 2010. IEEE Computer Society.

[6] Bahjat Safadi and Georges Quénot. Evaluations of multi-learner approaches for concept indexing in video documents. In *RIAO*, pages 88–91, 2010.

[7] Bahjat Safadi and Georges Quénot. Re-ranking by local re-scoring for video indexing and retrieval. In *CIKM*, pages 2081–2084, 2011.

[8] J. R. Smith, M. Naphade, and A. Natsev. Multimedia semantic indexing using model vectors. In *Proceedings of ICME - Volume 1*, pages 445–448, Washington, DC, USA, 2003. IEEE Computer Society.

[9] Ming-Fang Weng and Yung-Yu Chuang. Multi-cue fusion for semantic video indexing. In *Proceedings of the 16th ACM international conference on Multimedia*, MM '08, pages 71–80, New York, NY, USA, 2008. ACM.

[10] Y. Wu, Belle L. Tseng, and John R. Smith. Ontology-based multi-classification learning for video concept detection. In *ICME*, pages 1003–1006, 2004.

[11] Jun Yang and Alexander G. Hauptmann. (un)reliability of video concept detection. In *CIVR'08*, pages 85–94, 2008.

[12] Yingbin Zheng, Renzhong Wei, Hong Lu, and Xiangyang Xue. Semantic video indexing by fusing explicit and implicit context spaces. In *Proceedings of the international conference on Multimedia*, MM '10, pages 967–970, New York, NY, USA, 2010. ACM.